SRI KRISHNA COLLEGE OF ENGINEERING AND TECHNOLOGY

Cognizant Hackathon 2025

BATCH – 1

USE CASE NO – 3

USE CASE - PREDICTING MEDICAL EQUIPMENT FAILURE

1. PROBLEM STATEMENT

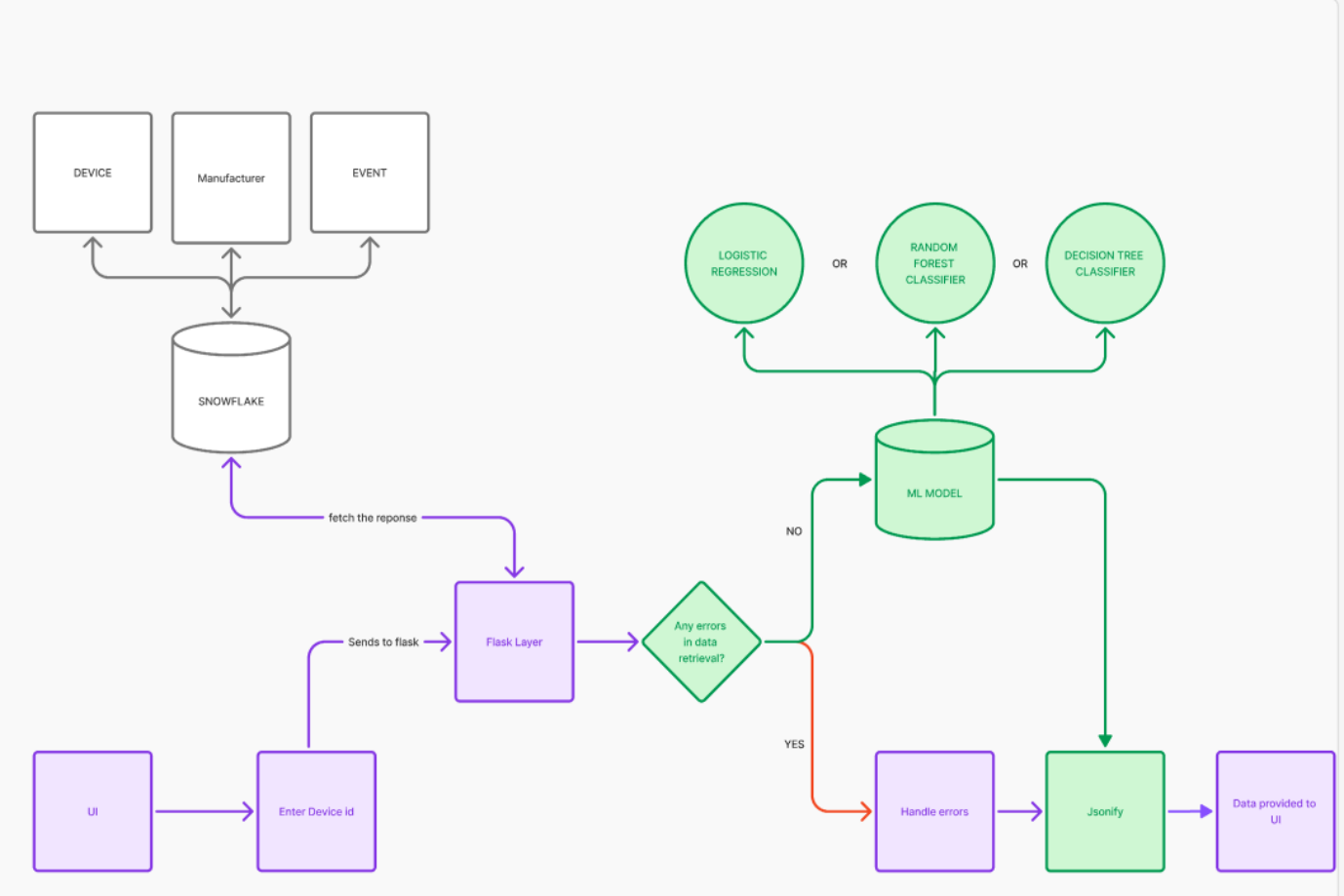
We aim to predict the likelihood and severity of faults in medical devices based on global recall, safety alert, and field notice data.

This is treated as a classification problem: identifying whether a device fault/recall is critical (within 50 days) or not.

2. DATASET SUMMARY

- **Source:** Faulty Medical Devices – Global Dataset (Kaggle)
- **Records:** ~120K entries across 3 CSV files (~36 MB)
- **Key Variables:** Device category, manufacturer, country, date issued, type of notice, and descriptive text.
- **Preprocessing:**
    1. Standardized dates and engineered recall duration
    2. Encoded categorical fields
    3. Handled missing values

3. FLOW DIAGRAM



3. MODELS TRIED

MODEL	PREPROCESSING	NOTES
Logistic Regression	One-hot Encoding and scaling	Fast baseline, interpretable
Decision Tree	Label encoding	Simple, explainable splits, but prone to overfitting
Random Forest	Label encoding	Strong non-linear capture,balanced

4. EVALUATION METRICS

- **Primary:** Accuracy, Precision, Recall, F1, ROC-AUC
- **Secondary:** Training time, interpretability, scalability
- **Rationale:** Balanced performance needed with hackathon feasibility.

5. RESULTS COMPARISON

MODEL	ACCURACY	F1	PRECISION	RECALL
Logistic Regression	70%	0.70	0.71	0.71
Decision Tree	72%	0.78	0.77	0.76
Random Forest	82%	0.80	0.82	0.80

1)LOGISTIC REGRESSION

```
Label distribution:
 label
0    57521
1    42479
Name: count, dtype: int64
Using features: ['id', 'action', 'action_classification', 'action_summary', 'reason', 'manufacturer_id', 'type', 'date_posted', 'status']
Accuracy: 0.7086

Classification Report:
              precision    recall  f1-score   support

           0       0.70      0.85      0.77     11504
           1       0.72      0.51      0.60      8496

    accuracy                           0.71     20000
   macro avg       0.71      0.68      0.68     20000
weighted avg       0.71      0.71      0.70     20000

{'device_id': 12530, 'failure_prediction': 1, 'risk_percentage': 77.27, 'within_50_days': 'Yes'}
```

2)DECISION TREE

```
✅ Accuracy: 0.72

Classification Report:

              precision    recall  f1-score   support

           0       0.72      0.75      0.70      8496
           1       0.70      0.72      0.70     11504

    accuracy                           0.70     20000
   macro avg       0.71      0.73      0.70     20000
weighted avg       0.72      0.73      0.70     20000


🔍 Prediction Example:

{'device_id': 19717, 'long_recall_prediction': 1, 'risk_percentage_long_recall': 60.45, 'recall_over_50_days': 'Yes'}
```

3)RANDOM FOREST

```
✅ Accuracy: 0.82

Classification Report:

              precision    recall  f1-score   support

           0       0.82      0.75      0.78      8496
           1       0.82      0.85      0.80     11504

    accuracy                           0.80     20000
   macro avg       0.82      0.80      0.79     20000
weighted avg       0.82      0.80      0.80     20000


🔍 Prediction Example:

{'device_id': 19717, 'long_recall_prediction': 1, 'risk_percentage_long_recall': 81.45, 'recall_over_50_days': 'Yes'}
```

6. OBSERVATIONS

- Logistic Regression → simple, interpretable, but misses patterns.
- Decision Tree → good all-around, interpretable, moderate time.
- Random Forest→ strongest overall performance (Accuracy 82%, PRECISION - 0.82).

7. FINAL MODEL SELECTION

Chosen Model: Random Forest

Why:

- Increased Performance – Random Forest provided consistently high accuracy and F1-scores, capturing non-linear relationships better than Logistic Regression while being less prone to overfitting.

- Interpretability – It offers straightforward feature importance scores, making it easier to explain results to stakeholders (important in healthcare contexts).

- Generalization – Handles categorical + numerical data effectively, robust to missing values and noisy features.

8. NEXT STEPS

Implementing XGBoost

- Experiment with models like XGBoost for increased accuracy and scalability and non-linear capture of patterns.

Feature Engineering

- Add domain-specific features (e.g., severity categories, region-based groupings, recall duration bins).

- Create interaction terms between device type and region.

Textual Data Handling (NLP)

- Apply TF-IDF or embeddings on recall reason/description text to incorporate richer signals into Random Forest.